Multi-step Jailbreaking Privacy Attacks on ChatGPT

Year: 2023 | Citations: 426 | Authors: Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, Jie Huang

---

Abstract

With the rapid progress of large language models (LLMs), many downstream NLP tasks can be well solved given appropriate prompts. Though model developers and researchers work hard on dialog safety to avoid generating harmful content from LLMs, it is still challenging to steer AI-generated content (AIGC) for the human good. As powerful LLMs are devouring existing text data from various domains (e.g., GPT-3 is trained on 45TB texts), it is natural to doubt whether the private information is included in the training data and what privacy threats can these LLMs and their downstream applications bring. In this paper, we study the privacy threats from OpenAI's ChatGPT and the New Bing enhanced by ChatGPT and show that application-integrated LLMs may cause new privacy threats. To this end, we conduct extensive experiments to support our claims and discuss LLMs' privacy implications.